Final Project Abstract

Chun Xue

The primary goal of this project is to develop a model to predict the box office of a new movie by using the data of previous movies. To reach this goal, there are three main steps required. The first step is data collecting, a scraper can be useful to collect the data. In this project, a scraper will be used to collect the features of movies like the cast, year, and film genre from related websites like IMDB and Wikipedia. There are also some existed datasets about movies, and the box office can be used as data source. The second step is cleaning and organizing data. Since the model is designed for predicting the box office of new movies, the taste of current audience should be considered. Therefore, the dataset which is used for training model should concentrate on movies released in the last five years, and very old movies should not be involved. The final dataset is a CSV file, and each row in this file represent the information (features) and box office (target) of a movie. The third step is training and testing the model. The most effective way to do this is to develop the program by using Python and some related packages like Pandas, Scikit-Learn, and TensorFlow. For this project, there are many different algorithms that can be used for building up the model, and each algorithm has advantages and disadvantages. In this project, three models will be trained independently by using three different algorithms. The first algorithm is Linear Regression, and the second is Classification and Regression Tree (CART), the third is Neural Networks. Since all these models are using the same training and testing dataset, by analyzing the test results, which algorithm has better performance on this project can be discussed and determined.

The idea of comparing the performance different models on a single task is a good one. The use of IMDB data to predict the box office total for a movie is *very* common in blog posts and other similar online sources. It will therefore be important to have an analysis which is very detailed, and distinct from previously done work.